Estimating Rate of Occurrence of Rare Events with Empirical Bayes: A Railway Application

John Quigley (*Corresponding author*), Tim Bedford and Lesley Walls

Department of Management Science

University of Strathclyde, Glasgow, G1 1QE, SCOTLAND

j.quigley@strath.ac.uk

Tel +44 141 548 3152

Fax +44 141 552 6686

Abstract

Classical approaches to estimating the rate of occurrence of events perform poorly when data are few. Maximum Likelihood Estimators result in overly optimistic point estimates of zero for situations where there have been no events. Alternative empirical based approaches have been proposed based on median estimators or non-informative prior distributions. While these alternatives offer an improvement over point estimates of zero, they can be overly conservative. Empirical Bayes procedures offer an unbiased approach through pooling data across different hazards to support stronger statistical inference.

This paper considers the application of Empirical Bayes to high consequence low frequency events, where estimates are required for risk mitigation decision support such as As Low As Reasonably Possible (ALARP). A summary of Empirical Bayes methods is given and the choices of estimation procedures to obtain interval estimates are discussed. The approaches illustrated within the case study are based on the estimation of the rate of occurrence of train derailments within the UK. The usefulness of Empirical Bayes within this context is discussed.

Key words: Empirical Bayes, Credibility Theory, Zero-failure, Estimation, Homogeneous Poisson Process

1. Introduction

The Safety Risk Model (SRM) is a large scale Fault and Event Tree model used to assess risk on the UK Railways. The objectives of the SRM (see Risk Profile Bulletin [1]) are to provide an understanding of the nature of the current risk on the mainline railway and risk information and profiles relating to the mainline railway. The main purpose of the model is to assist in the validation and development of Railway Safety Cases and support ALARP (As Low As Reasonably Possible) assessments. The model comprises 122 hazardous events and over 4000 end states. Developing reliable estimates for the rate of occurrence of rare events based on few data is a modern challenge facing the Railway Safety and Standards Board (RSSB), who are responsible for maintaining the SRM.

Classical approaches to estimating the rate of occurrence of events include calculating the ratio of the number of events that have occurred to the duration of the period of observation. In theory this procedure has desirable asymptotic properties, such as being an unbiased estimate of the rate of occurrence of such incidents, and being the Minimum Variance Unbiased Estimator. However it performs poorly when few data are available. For example, there are two obvious shortcomings with this approach. Firstly, there is a significantly large probability that the estimate will be 0, which is inappropriate for many situations, e.g. an event that occurs at a rate of 1 in 20 years has probability 0.6 of having an estimated rate of occurrence of 0 after 10 years of observation. Secondly, the variability of the estimate can be substantial from year to year, as in the previous example with probability 0.05 there would be at least an event per year, so the estimated rate of occurrence could switch from 0 events per annum after 10 years of observation to 0.091 events per annum in the eleventh year.

There are a number of different methods for estimating failure rates based on zero failure data. Direct assessment by an expert, such as advocated in a fully subjective approach, is one approach. Statistically driven approaches include the chi-squared method, which was developed in the early 1970's as a way of making a frequency parameter choice when there is no data [2]; the median estimate, where the estimated rate of occurrence is obtained through matching the probability of realising zero events during the observed time period with 0.5 [3]; or through assigning a uninformative prior distribution about the parameter of interest [3]. These methods may be suitable for estimating the rate of occurrence of events in isolation; however, we have a portfolio of events and for each we require an estimate.

The methodology we advocate is based on pooling data from various events to estimate an overall rate and then estimating appropriate adjustments from the pooled rate for each individual event. Such approaches to estimation appear under different guises, e.g. Credibility Theory or Empirical Bayes, but adhere to a common principle of estimating individual rates relative to an overall rate.

Credibility Theory began as an *ad hoc* method almost 100 years ago in the insurance industry. The basis of the method was to predict an individual's claim amount as a weighted average between the individuals past experience and the mean for individuals associated risk group. The link between this method and Empirical Bayes became apparent when the former was more rigorously developed in the 1960's [4]. For further discussion see [5].

Empirical Bayes (EB) methods are distinct from Bayesian methods as they use pooled data to estimate the prior distribution and therefore do not utilise subjective probability distributions. As such, construction of interval estimates is more complicated with EB because we must take account of the variability inherent in the occurrence of a particular event in addition to the estimation uncertainty associated with the construction of the empirical prior distribution.

The application of EB within the context of risk or reliability is not new. Martz and Waller [6] discuss the technique generally, [7] discuss an application to emergency diesel generators for binary data, [8, 9] use EB for estimating the rate of common cause failures, [10] use EB to support inference for the Weibull distribution within a software reliability growth context, [11] combine spline density estimates for prior distributions with EB for inference with the exponential distribution, [12] use EB within a Poisson modelling framework, and [13] makes use of the methodology with response surface modelling of categorical quality characteristics of possible designs.

A summary of the mechanics involved in EB procedures is presented in Section 2. We restrict ourselves to the Homogenous Poisson Process (HPP) and its conjugate prior the Gamma distribution. Point estimate procedures are provided together with discussion of the choices available for developing interval estimates.

An illustrative example is provided in Section 3 where we explore the application of EB for a subset of the events within the SRM. Point and interval estimates are provided, pool validation issues are discussed and the accuracy of the procedure is explored.

Finally, in Section 4 we reflect on the strengths and weaknesses of EB and propose further direction for future research.

2. Empirical Bayes – Background

2.1 The model

The prior distribution is used to describe the variability in the rate of occurrence within a pool of precursors, prior to observing any data. The prior is denoted by $\pi(\lambda)$, which is the probability density function measuring the likelihood of an event, chosen at random having a rate of occurrence of λ. We use the data for each specific precursor, i.e. hazardous event, to update the prior, refining the estimate uniquely for each precursor. The updated prior is referred to as the posterior distribution.

The SRM assumes that the rate of occurrence of events follows a constant rate over distance, which implies a Homogeneous Poisson Process. The distribution of N_i, the number of events that are realised when k_i miles are travelled, is Poisson.

$$P\left(N_i = n_i\right) = \frac{\left(\lambda_i k_i\right)^{n_i} e^{-\lambda_i k_i}}{n_i!}, \lambda_i > 0, k_i > 0, n_i = 0,1,.. \qquad (1)$$

An assumption about the parametric form of the prior distribution is made for numerical convenience. As we are assuming that the number of events follows a Poisson Process, a computationally convenient distribution to describe the prior distribution would be a Gamma distribution. This is a flexible distribution, which is able to describe a variety of different forms, defined by the following formula.

$$\pi\left(\lambda_1\right) = \frac{\beta^{\alpha} \lambda_1^{\alpha-1} \mathrm{e}^{-\beta\lambda}}{\Gamma\left(\alpha\right)}, \quad \lambda_1, \alpha, \beta > 0 \qquad (2)$$

The rate of occurrence for any particular event is not known, only that the rate has been selected at random from a Gamma distribution. We take an average of the Poisson distributions; weighted against the prior distribution. This provides the probability distribution of the number of events that will occur for precursor *i*, based only on our knowledge of the pool, i.e. the prior distribution. The following result due to Greenwood and Yule [14] (as cited by [15]) shows that the distribution of N_i is Negative Binomial.

$$\begin{aligned} P\left(N_i = n_i\right) &= \int_0^{\infty} \frac{\left(\lambda_i k_i\right)^{n_i} e^{-\lambda_i k_i}}{n_i!} \frac{\beta^{\alpha} \lambda_i^{\alpha-1} e^{-\beta\lambda_i}}{\Gamma(\alpha)} d\lambda \\ &= \frac{\Gamma\left(n_i+\alpha\right)}{\Gamma(\alpha) n_i!} \left(\frac{\beta}{\beta+k_i}\right)^{\alpha} \left(\frac{k_i}{\beta+k_i}\right)^{n_i}, \quad \alpha > 0, \beta > 0, n_i = 0,1,2,\ldots \end{aligned} \tag{3}$$

We treat the observed data as though they have been generated from this Negative Binomial distribution and estimate the parameters α and β from the data.

The following results concerning point and interval estimates assume a gamma prior distribution for the Homogeneous Poisson Process. The choice of parametric form for the prior should be verified for each specific application because it can have a significant impact on the estimates since it characterises the variability within the pool, from which we infer how much weight to place on an individual hazards experience compared with the pooled aggregate. Non parametric EB methods exist but the rate of convergence to the minimum Bayes risk will be much slower, see [6].

2.2 Point Estimates

Maximum Likelihood Estimates (MLE) can be pursued pooling all the data from all the events, using the Negative Binomial model. However, closed form MLE estimation equations do not exist for these parameters. If closed form equations are desired then the following approach can be used [16].

$$\hat{\alpha} = \frac{U^2}{W - U^2} \qquad \hat{\beta} = \frac{U}{W - U^2} \tag{4}$$

where:

$$U = \frac{\sum_{i=1}^{m} n_i}{\sum_{i=1}^{m} k_i} \tag{5}$$

which is the overall rate of occurrence of events and

$$W = \frac{\sum_{i=1}^{m} n_i^2 - \sum_{i=1}^{m} n_i}{\sum_{i=1}^{m} k_i^2} \tag{6}$$

which captures information about the second moment within the pool of events. These estimates are consistent estimators of the parameters (see Appendix). The MLE approach will provide more accurate estimates but suffer from requiring more computational effort.

Once an estimate of the prior distribution is obtained, Bayes Theorem is used to update the prior for each individual precursor to obtain the posterior distribution. The posterior distribution for the i^{th} event is:

$$\pi\left(\lambda_i \middle| N_i = n_i, \alpha, \beta\right) = \frac{\left(\beta + k_i\right)^{\alpha + n_i} \lambda_i^{\alpha + n_i - 1} e^{-(\beta + k_i)\lambda_i}}{\Gamma\left(\alpha + n_i\right)} \tag{7}$$

The Empirical Bayes estimate of λ_i is the mean of the posterior distribution. Therefore, the Empirical Bayes estimate of the rate of occurrence of event *i* is:

$$
\begin{aligned}
E\left(\lambda_i \middle| N_i = n_i\right) &= \int_0^\infty \lambda_i \pi\left(\lambda_i \middle| N_i = n_i, \alpha, \beta\right) d\lambda_i \\
&= \frac{\hat{\alpha} + n_i}{\hat{\beta} + k_i} \\
&= \frac{\hat{\alpha}}{\hat{\beta}}\left(1 - z\right) + \frac{n_i}{k_i} z
\end{aligned}
\tag{8}
$$

where:

$$
z = \frac{k_i}{\hat{\beta} + k_i} \tag{9}
$$

The Empirical Bayes estimate is a weighted average between the estimates from the pool, i.e. $\frac{\hat{\alpha}}{\hat{\beta}}$, and the traditional estimate of the individual precursor, i.e. $\frac{n_i}{k_i}$ *i*. As more data are obtained, k_i increases, and more weight is applied to the observed frequency.

2.3 *Interval Estimates*

The calculation of confidence intervals is not straightforward. If the true values of α and β were known then the posterior distribution could be used to assess the uncertainty in estimating λ_i. However, α and β have been estimated and as such we must account for the uncertainty in these estimates when developing true confidence intervals. Not accounting for the variability in the estimation of the posterior results in "naïve" intervals.

There are two sources of uncertainty within the estimation procedure. Firstly, there is the characteristic of the pool, i.e. the variability of frequencies for precursors that exist. Associated with this estimation will be sampling error. The second form of uncertainty is in the estimation of the rate of occurrence of incidence for a particular precursor. Naïve confidence intervals evaluate confidence intervals addressing only the second form of uncertainty, assuming the characteristics of the pool are known. If the pool consists of a very large number of precursors then the sampling error associated with estimating the parameters of the pool will be small and the naïve confidence intervals will be accurate. However, if sampling error is substantial then the naïve confidence intervals will under-estimate the true confidence intervals for each precursor.

There are different approaches available to constructing interval estimates for this model. We consider two approaches for assessing uncertainty with the prior distribution. The first approach developed by [17] assumes that the shape parameter, i.e. α, is known and the uncertainty is assessed through a distribution on the scale parameter, i.e. β. The second approach makes use of the limiting distribution of the Likelihood Ratio Statistic [18]. The latter approach does not rely on simplifying assumptions to the same extent as the former method but is more computationally demanding. Both approaches could suffer from double counting the data if the data from a particular hazard are used to assess uncertainty in the estimate conditioned on the prior distributions as in (7) and subsequently the data from all hazards are used to assess the uncertainty within the estimates of the prior distribution. This may be a problem if there are few hazards within the portfolio and as such the contribution from each hazard is significant. The influence of the data from each hazard on the uncertainty of the prior distribution can be assessed by estimating the uncertainty about the prior distribution using data from all hazards except the one being assessed for comparison. It is argued in [12] that this shortcoming is not present in the point estimates.

Other approaches exist such as using a non-informative improper joint prior distribution for (α, β) such as [19, 20] but have been criticised for resulting in improper posterior distributions [21] and could also suffer from double counting data.

2.3.1 Uniform Prior on β

Following this approach, we would use a non-informative prior distribution on β, which is to say that prior to observing any data we have no information about the true value of β. If we possessed prior information about the value of β then we could use more informative priors, of which one with a Gamma distribution form would be numerically convenient.

Updating the non-informative prior in light of the data provides the following posterior distribution.

$$\pi\left(\beta \middle| \underset{\sim}{n}, \alpha\right) = \frac{\prod_{i=1}^{m} \frac{\Gamma(n_i+\alpha)}{\Gamma(\alpha) n_i!} \left(\frac{\beta}{\beta+k_i}\right)^{\alpha} \left(\frac{k_i}{\beta+k_i}\right)^{n_i}}{\int_0^{\infty} \prod_{i=1}^{m} \frac{\Gamma(n_i+\alpha)}{\Gamma(\alpha) n_i!} \left(\frac{\beta}{\beta+k_i}\right)^{\alpha} \left(\frac{k_i}{\beta+k_i}\right)^{n_i} d\beta} \tag{10}$$

If exposure was the same for all events, such that $k_i = k$ then (10) would reduce to the following.

$$\pi\left(\beta \middle| \underset{\sim}{n}, \alpha\right) = \frac{\dfrac{\beta^{m\alpha}}{(\beta+k)^{m\left(\alpha+\bar{n}\right)}}}{\displaystyle\int_0^\infty \frac{\beta^{\alpha m}}{(\beta+k)^{m\left(\alpha+\bar{n}\right)}}\, d\beta}$$

$$= \frac{\Gamma\left(m\alpha + m\bar{n} + 4\right)}{\Gamma(m\alpha+1)\Gamma\left(m\bar{n}+3\right)} k \left(\frac{\beta}{\beta+k}\right)^{m\alpha} \left(\frac{k}{\beta+k}\right)^{m\bar{n}}$$

where:

$$\bar{n} = \frac{\sum_{i=1}^{m} n_i}{m}$$

which transforms to a Beta distribution with a change of variable.

$$\pi\left(p \middle| \underset{\sim}{n}, \alpha\right) = \frac{\Gamma\left(m\alpha + m\bar{n} + 4\right)}{\Gamma(m\alpha+1)\Gamma\left(m\bar{n}+3\right)} (1-p)^{m\alpha}\ p^{m\bar{n}+2}$$

$$where:$$

$$p = \frac{k}{k+\beta}$$

This distribution accounts for uncertainty in estimating β and as such would be used formally with the posterior distribution for λ_i to obtain confidence intervals.

If the true value of β were known, the posterior could be used to evaluate confidence intervals. While we do not know for certain the true value of β, we do have a probability distribution describing the likelihood of it taking particular values. Therefore, we wish to average the posterior distribution for λ_i across possible values of β weighted against the probability distribution (10).

$$\pi\left(\lambda_i \middle| N_i = n_i, \alpha, \underset{\sim}{n}\right) = \int_0^\infty \pi\left(\lambda_i \middle| N_i = n_i, \alpha, \beta\right) \pi\left(\beta \middle| \underset{\sim}{n}, \alpha\right) d\beta \qquad (11)$$

Mathematically, this is not tractable. However, it can be approximated by the following:

$$\pi\left(\lambda_i \middle| N_i = n_i, \alpha, \underset{\sim}{n}\right) = \frac{\sum_{j=i}^{J} \pi\left(\lambda_i \middle| \underset{\sim}{n}, \hat{\alpha}, \beta_j\right)}{J} \tag{12}$$

where β_j is the $100\left(\frac{j}{J+1}\right)^{th}$ percentile of the posterior distribution for β.

$$\int_0^{\beta_j} \pi\left(y \middle| \underset{\sim}{n}, \alpha\right) dy = \frac{j}{J+1} \tag{13}$$

The choice of J depends on how accurate we wish to be. The larger the value of J the greater the accuracy but the more calculations are required.

Having approximated the posterior distribution, we solve the appropriate percentiles determined by the level of confidence we seek.

One shortcoming with this approach lies in the assumption that α is known. The MLE of α and β are highly correlated hence assuming the value of α is known severely restricts the variability in β.

2.1.2 Likelihood Ratio Statistic

The limiting distribution of –2 times the natural logarithm of the relative likelihood function has a χ^2 distribution with 2 degrees of freedom. This is expressed in (14) where $\left(\hat{\alpha}, \hat{\beta}\right)$ are the MLE. This can be used to construct a joint confidence region for the parameters.

$$-2\ln\left(\frac{\prod_{i=1}^{m}\frac{\Gamma(n_i+\alpha)}{\Gamma(\alpha)n_i!}\left(\frac{\beta}{\beta+k_i}\right)^{\alpha}\left(\frac{k_i}{\beta+k_i}\right)^{n_i}}{\prod_{i=1}^{m}\frac{\Gamma\left(n_i+\hat{\alpha}\right)}{\Gamma\left(\hat{\alpha}\right)n_i!}\left(\frac{\hat{\beta}}{\hat{\beta}+k_i}\right)^{\hat{\alpha}}\left(\frac{k_i}{\hat{\beta}+k_i}\right)^{n_i}}\right)\sim\chi^2_{(2)} \qquad (14)$$

To construct a tolerance interval about (7) we first determine the locus of point points for α and β such that the Cumulative Distribution Function (CDF) of (14) does not exceed a specified value assuming only one degree of freedom rather than two. For further discussion see [18].

There is no closed form solution for this approach to obtaining confidence intervals and computationally it can be intense. However, we formally make use of the correlation between the estimates within this approach.

3. Case Study – SRM

The application of the methods described in Section 2 is illustrated using a subset of data from the SRM, namely the passenger derailment events. There are 59 possible events and there have been 66 relevant occurrences in the past 6 years providing an overall empirical estimate of the rate of 6.6E-10 events per mile. 54% of the 59 events have had no realisations in the past 6 years.

The analysis protocol is as follows. First the prior distribution for the rate of occurrence of events for the group is determined and goodness-of-fit of a Gamma distribution as its parametric form is assessed. Second, point estimates for the rate of occurrence of events for each precursor are derived. Third confidence intervals for the parameters are obtained using both methods described in Section 2 for comparison. Fourth, the impact of partitioning the precursors into sub-groups and reassessing the rates is investigated. Finally, we reflect upon the difference between the rates and the estimates currently used with the SRM.

3.1 Empirical Prior Distribution for Number of Events

Treating the number of occurrences as though they were realized from a Negative Binomial distribution, the MLE of α and β are:

$$\hat{\alpha} = 0.43$$
$$\hat{\beta} = 5.95\text{E+8} \tag{15}$$

While the MLE's will be more accurate, they require solving equations involving the gamma function. We calculated the point estimate using the closed form solutions in (4) for comparison, which resulted in (16).

$$\hat{\alpha} = 0.55$$
$$\hat{\beta} = 8.32\text{E} + 8 \tag{16}$$

The MLE's in (15) produce an overall average rate of occurrence within the pool of 7.22E-10 events per mile, which is less than 10% greater than the estimate obtained when using (16). Figure 1 is an illustration of the empirical distribution compared with the Negative Binomial where the parameters were estimated through MLE (15).

Figure 1 indicates a reasonably close fit between the data and the model. Formally a χ^2 goodness-of-fit test where the null model is a Negative Binomial with parameter values as in (15) was not rejected at the 1% significance level.

INSERT FIGURE 1

3.2 Point Estimates for Event

The traditional estimate for each precursor is to calculate the ratio between the number of events and the total number of train miles travelled. Using the MLE for α and β in (15) we propose the weight (9), i.e. z, applied to the traditional estimate to be 0.72 for those precursors that have had an exposure of 2.57E+8 miles per annum and a weight of 0.76 for those precursors that have had an exposure of 3.1E+8 miles per annum.

The EB estimate of the precursors and their observed frequencies, which are the traditional estimates, are illustrated in Figure 2. The data are summarised into categories depending on the number of events that have been realised by each precursor.

A phenomenon known as shrinkage is illustrated in Figure 2, whereby, the observed frequencies are drawn towards the pooled mean. The high frequency precursors have a lower estimate and the estimates for the precursors that are infrequent are increased towards the mean.

INSERT FIGURE 2

3.3 Interval Estimates

For comparison purposes we calculate the confidence interval for the shape parameter using the uniform prior approach as well as developing confidence intervals for the rate of occurrence of each event using the Likelihood Ratio approach. This results in the following posterior distribution for β and is illustrated in Figure 3.

$$\pi\left(\beta \middle| \underset{\sim}{n}, \alpha\right) = \frac{\dfrac{\beta^{25.37}}{(\beta + 1.54\text{E+9})^{54.33}(\beta + 1.86\text{E+9})^{37.04}}}{\displaystyle\int_0^\infty \frac{\beta^{25.37}}{(\beta + 1.54\text{E+9})^{54.33}(1.86\text{E+9})^{37.04}}\, d\beta} \tag{17}$$

The values of β range from approximately 0.4E+9 to 1.2E+9. Compare this with the joint 95% confidence region displayed in Figure 4 where β ranges from 0.3E+9 to 2.5E+9. As discussed in section 2 the uniform prior approach over-estimates the confidence provided by the data with a narrower confidence interval due to the assumption α is known.

INSERT FIGURE 3

INSERT FIGURE 4

3.4 Sensitivity of Pool Choice

The estimates for the rates of occurrence are re-assessed based on smaller pools. The events under consideration have one of two different annual miles exposure. An obvious partitioning of these events would be based on the different rates of exposure. We follow the same procedures used to evaluate the entire set, but for the two subsets.

The MLE's for the events with the annual exposure of 2.57E+8 miles per annum are:

$$\hat{\alpha} = 1.06$$
$$\hat{\beta} = 1.36\text{E} + 9$$

and for the events with annual exposure of 3.1E+8 miles per annum are:

$$\hat{\alpha} = 0.25$$
$$\hat{\beta} = 0.5\text{E+9}$$

Both MLE's are contained within the joint confidence interval of Figure 4. The resulting difference in point estimates for the rate of occurrence of each of the events has not been greatly affected. The arithmetic difference in the point estimates was negative for half the events and positive for the other half. The maximum difference was 5.75E-10, the smallest –1.64E-10 and 70% of the events had a different with order of magnitude E-11.

One effect of partitioning the pool was to alter the weight applied to the individual experience. The group with the annual exposure of 3.10E+8 miles moved from 76% to 79%. The reverse effect has occurred for the group with the smaller annual exposure of 2.57E+8 miles moving the weight from 72% to 53%.

3.5 Comparison of Results with the SRM

Estimates existed for these events prior to conducting this study. These are based on past experience and expert judgment. There was much agreement between the SRM and the Empirical Bayes estimates. The EB estimate resulted in a higher frequency for 32 of the 59 events. The minimum arithmetic difference was –1.75E-10, where a negative number indicates the EB estimated a higher frequency and the maximum difference was 1.65E-8. The mean difference was 4.4E-10 and only two events had a difference with order of magnitude greater than 10^{-10}. This methodology can be utilised to highlight those precursors where there is substantial difference for further investigation.

3.6 Testing Pools

The SRM is partitioned into a number of subsets of events based type of incidents being considered. We have been studying one particular partition, namely the passenger derailment events. However, these subsets are not mutually exclusive. As such we would have a dilemma as to which group to assign a particular event. We propose constructing a test to assess whether there is strong statistical evidence in favour of one pool, for a given event.

For each pool we obtain an estimate of the set of parameters, i.e. α and β. We consider which one of two pools, say the null and alternative pools, to assign a precursor. We have two probabilities measuring the likelihood of the observed number of events being observed from the precursor, the null and alterative. We construct a test based on the ratio of these probabilities, the alternative divided by the null. If the ratio is large then more evidence exists in support of the alternative compared with a low ratio. We can define a critical value such that if the ratio is above the value then we do not have enough evidence to strongly support the null model.

Consider the event having the greatest change in point estimate between the larger pool and the smaller pool, as discussed in the previous section. This event belongs to the set of events with the smaller of the two exposures. Essentially we wish to test whether the data for this event is better modelled with a Negative Binomial distribution with parameter values of α=0.43 and β=5.95E+08, or with parameters α=1.06 and β=1.36E+09.

The bigger pooled group, i.e. using all the events, has been selected as the null model and the smaller pooled group as the alternative. For this particular event there were 6 realizations. The ratio of the probability of obtaining 6 realizations using the estimates obtained through the alternate model to the probability of obtaining 6 realizations using the null model is 0.83. Therefore, the evidence is on the side of the null model. Assuming the null model to be the correct model, the probability of obtaining a ratio of 0.83 or less is 0.62. Therefore it is not an unusual observation and we do not reject the null model, i.e. the bigger pool.

Reversing the labelling such that the smaller pool was the null model and the bigger pool was the alternative, we conclude that there was no strong evidence against the null model, i.e. the smaller pool, and as such there is no strong statistical evidence in favour of either pool.

3.7 Summary

The aim of this investigation was to determine a methodology that could be used to estimate the frequency of rare events where event data are not likely to exist. We have proposed an Empirical Bayes approach and illustrated its use with data.

The assumptions made in developing the model are applicable to any of the events in the SRM which assumes the rate of occurrence is constant. The EB estimates were in close agreement with the existing SRM estimates, although there were some notable differences. While there may exist sound reasons for not relying on the data for each event, this process will help identify estimates for closer inspection.

4. Conclusions

Point estimates obtained from Empirical Bayes procedures have been shown to fare no worse than traditional methods with respect to accuracy when the pool is very heterogeneous and better than traditional methods when the pool is homogeneous [22]. This is not surprising as the more heterogeneous the pool, the more weight is applied to the individual observations. Moreover, the approach of utilizing the data to estimate the prior distribution has been shown to be robust against prior misspecification (see [23]). Several case studies and investigations have been published illustrating the effectiveness of this methodology, see [22, 24, 25] for examples.

This approach has focused entirely on the data. However, a full Bayesian approach would utilise expert judgment. This could easily be integrated within the framework present. If we had prior knowledge of the relative difference in frequencies between events, rather than assuming they all are selected at random from the same pool, we could enhance the inference, and ensure that the overall subjective assessments were consistent with the data observed.

The greatest challenge for using this approach in practice is with the definition of pool membership. While in the example provided there was little difference between the frequency estimates in the two pools considered, the potential exists to greatly influence an estimated rate by assigning in to one pool or another. This method offers a great deal in improved accuracy of estimates with few data, but could be subject to abuse in this regard. As an illustration, through adding exceedingly rare hazardous events that have yet to be experienced to the pool, you would be decreasing the overall mean and therefore pull down the estimate of the rate of occurrence of each hazardous event. We feel there would be benefits from using structured expert judgement to assign hazardous events to pools, justified on non-empirical grounds and then use Empirical Bayes methods for estimation. We continue to research this issue.

Appendix

We will demonstrate that the Empirical Bayes estimator obtained using (6) and (7) is a consistent estimator.

The model describing the variability in the number of incidents is a Negative Binomial distribution.

$$P\left(N_i = n_i\right) = \frac{\Gamma\left(n_i + \alpha\right)}{\Gamma\left(\alpha\right) n_i!}\left(\frac{\beta}{\beta + k_i}\right)^{\alpha}\left(\frac{k_i}{\beta + k_i}\right)^{n_i}$$

where k_i is the measure of exposure.

The expectation and variance of the Negative Binomial distribution are:

$$E\left[N_i\right] = k_i \frac{\alpha}{\beta}$$

$$Var\left[N_i\right] = k_i \frac{\alpha}{\beta}\left(\frac{k_i + \beta}{\beta}\right)$$

Consider the following.

$$U = \frac{\sum_{i=1}^{m} N_i}{\sum_{i=1}^{m} k_i}$$

The expectation of *U* is:

$$E\left[U\right]=E\left[\frac{\sum_{i=1}^{m}N_i}{\sum_{i=1}^{m}k_i}\right]$$

$$=\frac{E\left[\sum_{i=1}^{m}N_i\right]}{\sum_{i=1}^{m}k_i}$$

$$=\frac{\sum_{i=1}^{m}E\left[N_i\right]}{\sum_{i=1}^{m}k_i}$$

$$=\frac{\alpha}{\beta}$$

The variance of U is:

$$Var\left[U\right]=Var\left[\frac{\sum_{i=1}^{m}N_i}{\sum_{i=1}^{m}k_i}\right]$$

$$=\frac{\sum_{i=1}^{m}Var\left[N_i\right]}{\left(\sum_{i=1}^{m}k_i\right)^2}$$

Substituting the variance of a Negative Binomial into this formula we have the following.

$$Var\left[U\right]=\frac{\frac{\alpha}{\beta}\sum_{i=1}^{m}\left(\frac{k_i}{\beta}+1\right)k_i}{\left(\sum_{i=1}^{m}k_i\right)^2}$$

$$=\frac{\alpha}{\beta}\left(\frac{1}{\left(\sum_{i=1}^{m}k_i\right)}+\frac{\sum_{i=1}^{m}k_i^2}{\beta\left(\sum_{i=1}^{m}k_i\right)^2}\right)$$

As exposure, i.e. k_i, increases the variance of U decreases converging with its mean and therefore is a consistent estimator of α/β.

Consider

$$W = \frac{\sum_{i=1}^{m} n_i^2 - \sum_{i=1}^{m} n_i}{\sum_{i=1}^{m} k_i^2}$$

The expectation of W is:

$$\begin{aligned}
E[W] &= \frac{\sum_{i=1}^{m} E\left[N_i^2\right] - \sum_{i=1}^{m} E\left[N_i\right]}{\sum_{i=1}^{m} k_i^2} \\
&= \frac{\sum_{i=1}^{m}\left(\frac{\alpha k_i}{\beta}\right)\left(1+\frac{(1+\alpha)}{\beta}k_i\right) - \frac{\alpha}{\beta}\sum_{i=1}^{m} k_i}{\sum_{i=1}^{m} k_i^2} \\
&= \frac{\frac{\alpha(1+\alpha)}{\beta^2}\sum_{i=1}^{m} k_i^2}{\sum_{i=1}^{m} k_i^2} \\
&= \frac{\alpha(1+\alpha)}{\beta^2}
\end{aligned}$$

The variance of W is

$$Var[W] = \frac{E\left[\left(\sum_{i=1}^{m} N_i^2 - \sum_{i=1}^{m} N_i\right)^2\right]}{\left(\sum_{i=1}^{m} k_i^2\right)^2} - \left[\frac{\alpha(1+\alpha)}{\beta^2}\right]^2$$

$$= \frac{E\left(\sum_{i=1}^{m} N_i^2\right)^2 - 2\sum_{i=1}^{m} E\left[N_i^3\right] + E\left(\sum_{i=1}^{m} N_i\right)^2}{\left(\sum_{i=1}^{m} k_i^2\right)^2} - \left[\frac{\alpha(1+\alpha)}{\beta^2}\right]^2$$

$$= \frac{\sum_{i=1}^{m}\sum_{j=1}^{m} E\left[N_i^2 N_j^2\right] - 2\sum_{i=1}^{m} E\left[N_i^3\right] + \sum_{i=1}^{m}\sum_{j=1}^{m} E\left[N_i N_j\right]}{\left(\sum_{i=1}^{m} k_i^2\right)^2} - \left[\frac{\alpha(1+\alpha)}{\beta^2}\right]^2$$

$$= \frac{\begin{array}{l} \sum_{i=1}^{m}\sum_{\substack{j=1\\ j\neq i}}^{m}\left[\frac{\alpha}{\beta}k_i\left(1+\frac{\alpha+1}{\beta}k_i\right)\frac{\alpha}{\beta}k_j\left(1+\frac{\alpha+1}{\beta}k_j\right)\right] \\ +\sum_{i=1}^{m}\left[\frac{\alpha}{\beta}k_i + 7k_i^2\frac{\alpha(\alpha+1)}{\beta^2} + 6k_i^3\frac{\alpha(\alpha+1)(\alpha+2)}{\beta^3} + k_i^4\frac{\alpha(\alpha+1)(\alpha+2)(\alpha+3)}{\beta^4}\right] \\ -2\sum_{i=1}^{m}\left[\frac{\alpha}{\beta}k_i + 2k_i^2\frac{\alpha(\alpha+1)}{\beta^2} + k_i^3\frac{\alpha(\alpha+1)(\alpha+2)}{\beta^3}\right] \\ +\sum_{i=1}^{m}\sum_{\substack{j=1\\ j\neq i}}^{m}\frac{\alpha}{\beta}k_i\frac{\alpha}{\beta}k_j + \sum_{i=1}^{m}k_i^2\frac{\alpha(\alpha+1)}{\beta^2} \end{array}}{\left(\sum_{i=1}^{m} k_i^2\right)^2} - \left[\frac{\alpha(1+\alpha)}{\beta^2}\right]^2$$

A little manipulation results in the following.

$$Var[W] = \frac{\begin{array}{l} 2\left(\frac{\alpha}{\beta}\right)^2\left(\sum_{i=1}^{m}k_i\right)^2 - \frac{\alpha}{\beta}\sum_{i=1}^{m}k_i - \frac{\alpha^4+\alpha^2-4\alpha}{\beta^2}\sum_{i=1}^{m}k_i^2 + 2\left(\frac{\alpha}{\beta}\right)^2\left(\frac{\alpha+1}{\beta}\right)\sum_{i=1}^{m}\sum_{j=1}^{m}k_i k_j^2 \\ +\frac{8(\alpha+1)\alpha(1-\alpha)}{\beta^3}\sum_{i=1}^{m}k_i^3 + \frac{\alpha(\alpha+1)(\alpha+2)(\alpha+3)}{\beta^4}\sum_{i=1}^{m}k_i^4 \end{array}}{\left(\sum_{i=1}^{m}k_i^2\right)^2}$$

As exposure increases the variance decreases to 0 and therefore *W* is a consistent estimator of $\alpha(\alpha+1)/\beta^2$.

By Slutskys Theorem [26] any continuous mapping of estimators will converge to the continuous mapping of the asymptotic values of the estimators. Substituting the expectations into the estimation formulas results in the following.

$$\begin{aligned}
\lim_{k\to\infty}\hat{\alpha} &= \lim_{k\to\infty}\frac{U^2}{W-U^2} \\
&= \frac{\left(\frac{\alpha}{\beta}\right)^2}{\frac{\alpha(1+\alpha)}{\beta^2}-\left(\frac{\alpha}{\beta}\right)^2} \\
&= \alpha
\end{aligned}$$

$$\begin{aligned}
\lim_{k\to\infty}\beta &= \lim_{k\to\infty}\frac{U}{W-U^2} \\
&= \frac{\frac{\alpha}{\beta}}{\frac{\alpha(1+\alpha)}{\beta^2}-\left(\frac{\alpha}{\beta}\right)^2} \\
&= \beta
\end{aligned}$$

Therefore, (6) and (7) result in consistent estimators of the parameters. Moreover, the estimator of the rate of occurrence of events for any event will converge in distribution to the following.

$$\frac{\hat{\alpha}+\mathrm{n_i}}{\hat{\beta}+\mathrm{k_i}} \xrightarrow{d} \frac{\alpha+\mathrm{n_i}}{\beta+\mathrm{k_i}}$$

Acknowledgement

The authors would like to thank the UK Railway Safety and Standard Board (RSSB) for their support. This paper is based on work that was carried out for the RSSB [1].

References

1. www.rssb.co.uk
2. E Welker and M Lipow (1974), Estimating the exponential failure rate from data with no failure events, in *Proceedings of RAMS 1974,* IEEE Press.
3. R Bailey (1997) Estimation from Zero-Failure Data, *Risk Analysis*, **17**, pp 375-380.
4. H. Buhlmann (1967) Experience Rating and Credibility, *ASTIN Bulletin,* **4**, pp 199-207
5. S Klugman, H Panjer and G Willmot (1997) *Loss Models: From Data to Decisions*. John Wiley and Sons.
6. H. Martz and R Waller (1991) *Bayesian Reliability Analysis*, Kreiger Publishing Company.
7. WE Vesely, SP Uryasev and PK Samanta (1994) Failure of Emergency Diesel Generators: A Population Analysis Using Empirical Bayes Methods, *Reliability Engineering and System Safety*, **46**, pp 221-229.
8. JK Vaurio (2005) Uncertainties and Quantification of Common Cause Failure Rates and Probabilities for System Analyses, *Reliability Engineering and System Safety*, **90**, pp 186-195.
9. Vaurio JK (2002) Extesnions of the Uncertainty Quantification of Common Cause Failure Rates, **78**, pp 63-69.
10. J Ferdous, M Borhan Uddin and M Pandey (1995) Reliability Estimation with Weibull Inter Failure Times, *Reliability Engineering and System Safety*, **50**, pp 285-296.

11. F Grabski and A Sarhan (1996) Empirical Bayes Estimation in the Case of Exponential Reliability, *Reliability Engineering and System Safety*, **53**, pp 105-113.

12. JK Vaurio and KE Jankala (2006) Evaluatin and Comparison of Estimation Methods for Failure Rates and Probabilities, *Reliability Engineering and System Safety*, **91**, pp 209-221.

13. SY Sohn (1999) Robust Parameter Design for Integrated Circuit Fabrication Procedure with Respect to Categorical Characteristic, *Reliability Engineering and System Safety*, **66**, pp 253-260.

14. M Greenwood and GU Yule (1920) An inquiry into the nature of frequency distributions representative of multiple happening with particular reference to the occurrence of multiple attacks of disease or of repeated accidents, *Journal of the Royal Statistical Society A*, **83**, pp 255-279.

15. N. Johnson, S Kotz and A Kemp (1993) *Univariate Discrete Distributions 2nd Edition*, John Wiley and Sons

16. S. Arnold (1990) *Mathematical Statistics*, New Jersey, Prentice-Hall

17. J Deeley and D Lindley (1981) Bayes Empirical Bayes, *JASA*, **76**, pp 833-841.

18. J Lawless (2003) *Statistical Models and Methods for Lifetime Data 2nd Edition*, John Wiley and Sons.

19. S Hora and R Iman (1987) Bayesian analysis of learning in risk analysis, *Technometrics*, **29**, pp 221-228.

20. K Pörn (1990) *On Empirical Bayesian Inference Applied to Poisson Probability Models*, Linköping Studies in Science and Technology, Dissertation No 234, Linköping.

21. C Bunea (2003) *Mathematical Models for Reliability Data*, Doctoral disseration TU Deft Netherlands.

22. B Carlin and T Louis (2000) *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd Edition, Chapman and Hall/CRC.

23. C. Morris (1983) Natural exponential families with quadratic variance functions: Statistical theory, *Annals of Statistics*, **11**, pp 515-529

24. T Tomberlin (1988) Predicting Accident Frequencies for Drivers Classified by Two Factors, *JASA*, **83**, 309-321.

25. J Maritz and T Lwin (1989) *Empirical Bayes Methods*, 2nd Edition, London, Chapman and Hall.

26. K Knight (2000) *Mathematical Statistics,* Chapman and Hall.

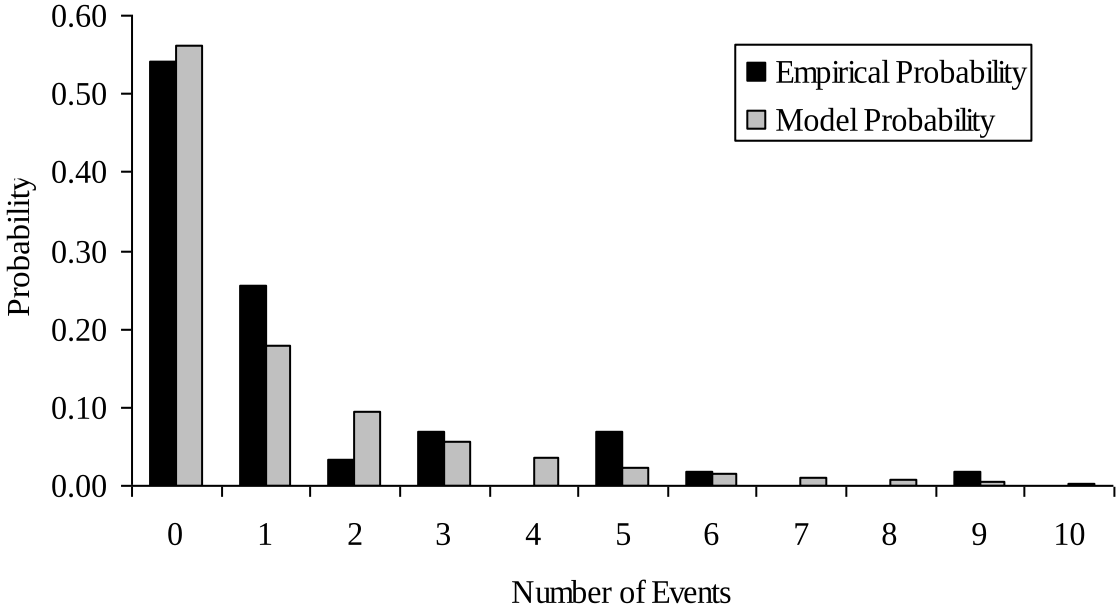
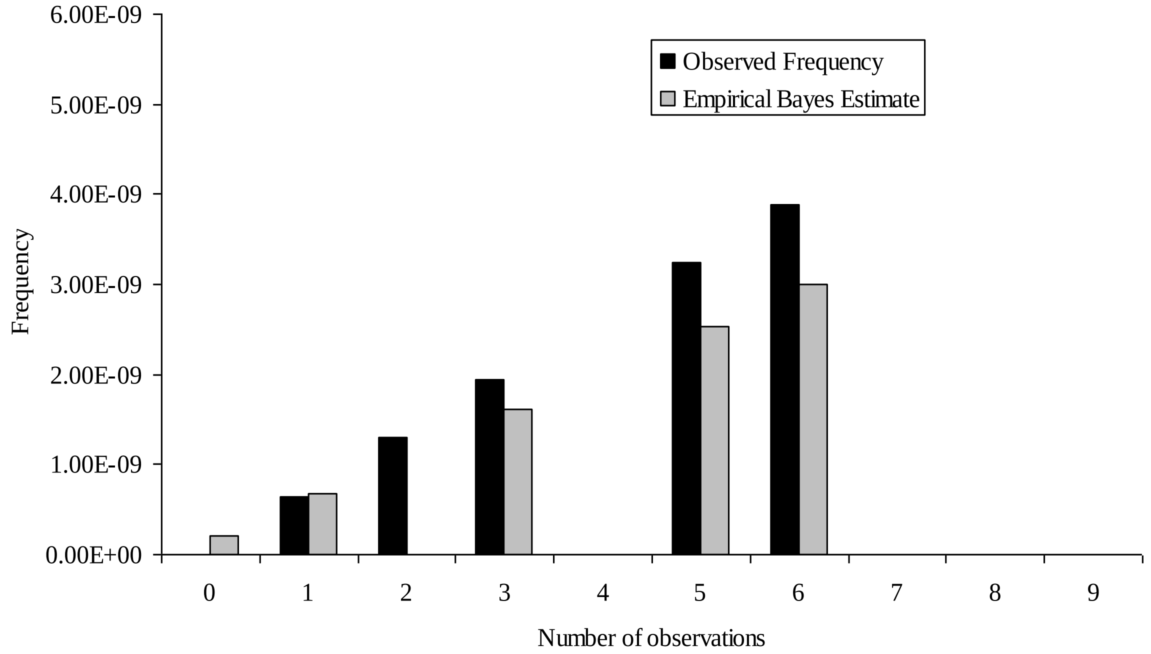
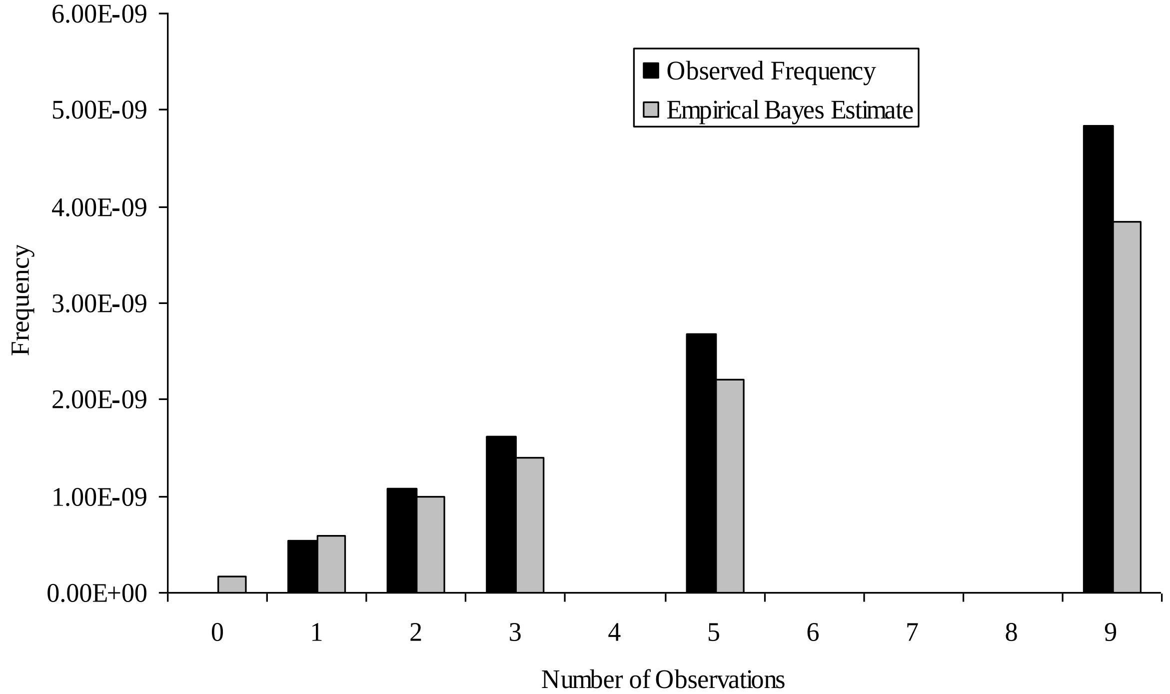


Figure 1 Comparison between empirical probability and model estimate



a) For events with exposure of 1.54E+9



b) For events with exposure of 1.86E+9

Figure 2 Comparison between observed and estimated frequency of occurrences

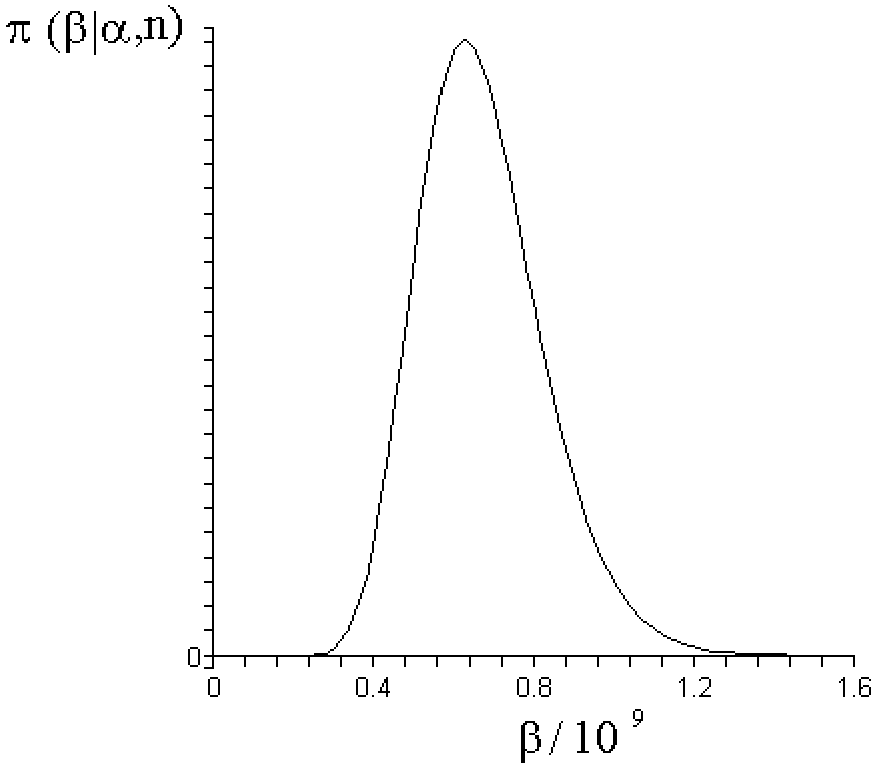


Figure 3 Posterior distribution used for scale parameter

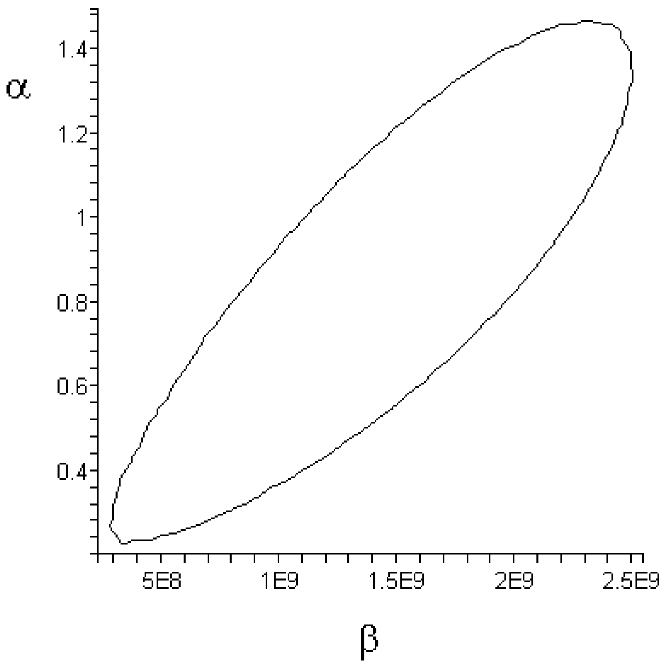


Figure 4 Joint 95% Confidence Region of α and β